Exploring the Factors Affecting Mental Health in the Tech Industry: A Machine Learning Analysis

Milestone - 4

Group 21

Student 1 - Aishwarya SVS

Student 2 - Mohit Chodisetti

857-376-1986

857-867-1576

svs.a@northeastern.edu

chodisetti.m@northeastern.edu

Percentage of Effort Contributed by Student 1: 50%

Percentage of Effort Contributed by Student 2: 50%

Signature of Student 1: Aishwarya SVS

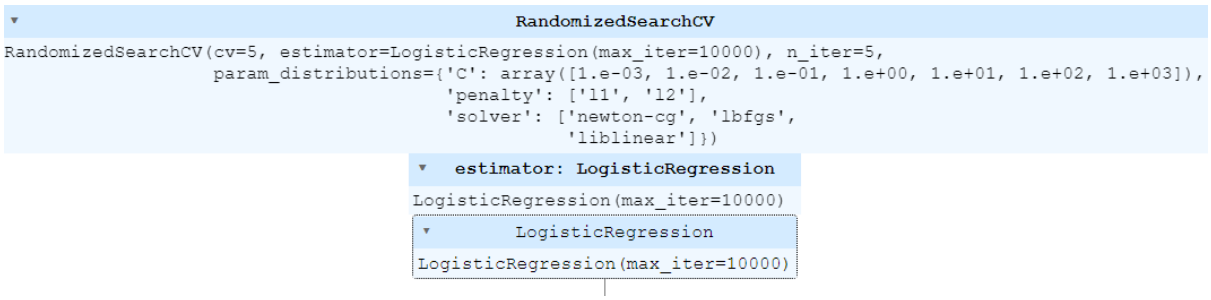
Signature of Student 2: Mohit.Chvk

Submission Date: 03-24-2023

Introduction:

This report presents the data mining models along with their respective Evaluation Metrics applied on the OSMI dataset used in the project, providing insights into mental health in the workplace in 2016. The dataset contains 1433 rows and 63 columns.

Logistic Regression:




```
Best hyperparameters: {'solver': 'newton-cg', 'penalty': 'l2', 'C': 1.0}
```

The logistic regression model was trained using a grid search with three hyperparameters: penalty ['l1','l2'], C np.logspace(-3,3,7), and solver['newton-cg', 'lbfgs', 'liblinear'] . The confusion matrix shows that the model was able to correctly classify a majority of the instances, with an accuracy of 76.31%. The recall score of 90.69% indicates that the model was able to correctly identify a high percentage of the positive instances, while the precision score of 76.46% indicates that it correctly identified a lower percentage of the positive instances. The F1 score of 82.97% indicates a good balance between precision and recall. The ROC AUC score of 79.17% indicates that the model can distinguish between the two classes moderately well. Overall, the logistic regression model with the chosen hyperparameters is moderately successful in classifying the instances in the OSMI dataset.

Also, the model achieved an accuracy score of 0.7368 on the test set. The confusion matrix shows that the model correctly predicted 23 out of 60 samples in the first class, 66 out of 79 samples in the second class, and 93 out of 108 samples in the third class. The model had the highest recall score of 0.8611 for the second class, indicating that it correctly identified a high proportion of the actual positive samples in that class. The model had the highest precision score of 0.7209 for the first class, indicating that of all the samples predicted to be positive in that class, the majority were true positives. The F1 score was 0.7848, which is the harmonic mean of precision and recall. The ROC AUC score was 0.7753, indicating that the model performed relatively well in distinguishing between positive and negative samples.

****** For Training on Logistic Regression Model ******

The Confusion Matrix:
 [[91 53 69]
 [25 283 48]
 [16 23 380]]

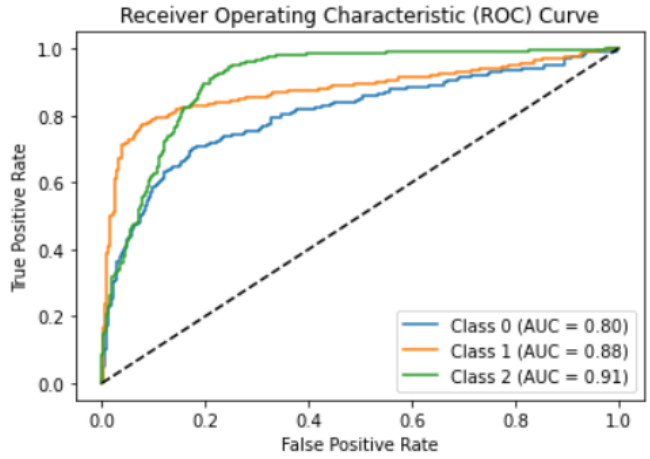
 The Accuray Score: 0.7631578947368421

 The Recall Score: 0.9069212410501193

 The Precision Score: 0.7645875251509054

 The F1 Score: 0.8296943231441049

 The ROC AUC Score: 0.7917191360463433



****** For Testing on Logistic Regression Model ******

The Confusion Matrix:
 [[23 14 23]
 [0 66 13]
 [6 9 93]]

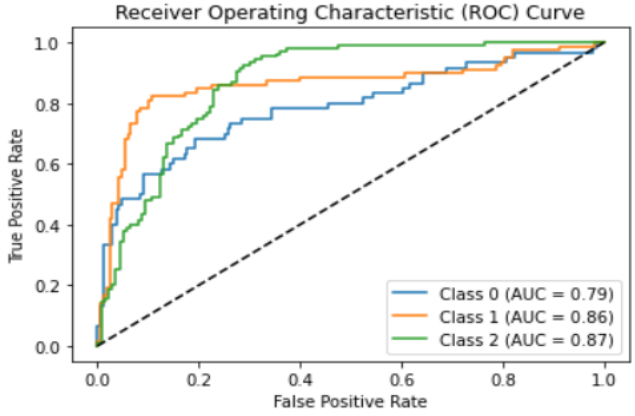
 The Accuray Score: 0.7368421052631579

 The Recall Score: 0.8611111111111112

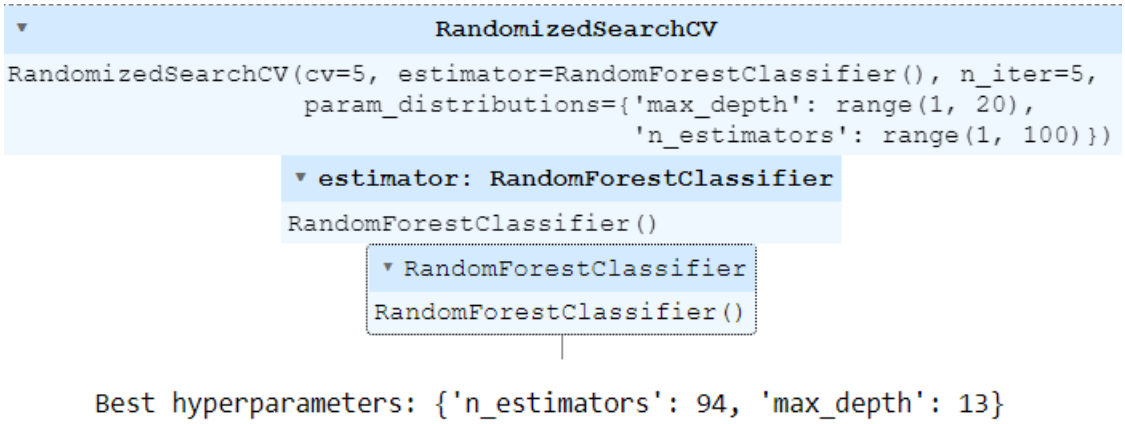
 The Precision Score: 0.7209302325581395

 The F1 Score: 0.7848101265822784

 The ROC AUC Score: 0.775317392210274



Random Forest:



Best hyperparameters: {'n_estimators': 94, 'max_depth': 13}

The Random Forest model was trained using a parameter grid search with n_estimators ranging from 1 to 100 and max_depth ranging from 1 to 20. The confusion matrix shows that the model correctly classified 131 instances of the first class, 308 instances of the second class, and 412 instances of the third class. However, there were also misclassifications, as shown in the confusion matrix.

The accuracy score of the model was 0.861, indicating that it correctly classified 86.1% of the instances in the dataset. The recall score, or the ability of the model to correctly identify true positives, was 0.983, indicating that the model correctly identified 98.3% of instances of all three classes. The precision score, or the ability of the model to correctly identify positive instances, was 0.798, indicating that the model correctly identified 79.8% of positive instances.
The F1 score, which is the harmonic mean of precision and recall, was 0.881, indicating that the model achieved a balance between precision and recall. Finally, the ROC AUC score was 0.872, indicating that the model had a good ability to distinguish between the three classes.

Also, the model achieved an accuracy of 0.77 on the test data. It correctly identified 96% of the positive cases (high recall), but its precision was relatively low at 72%. The F1 score was 0.83, indicating a reasonably good balance between precision and recall. The ROC AUC score was 0.80, which indicates that the model is better than random at distinguishing between positive and negative cases.

****** For Training on Random Forest Model ******

The Confusion Matrix:
 [[131 21 61]
 [5 308 43]
 [4   3 412]]

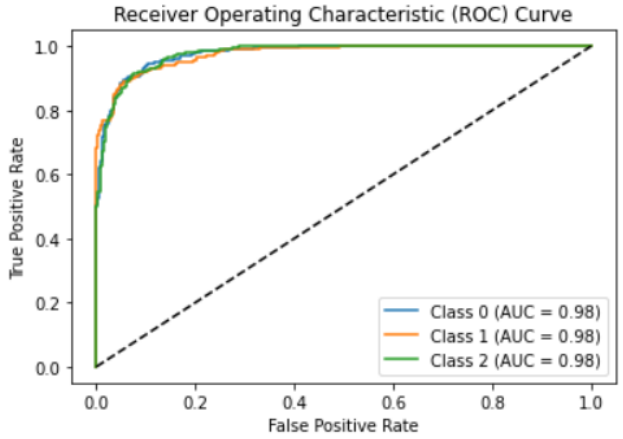
 The Accuray Score: 0.861336032388664

 The Recall Score: 0.9832935560859188

 The Precision Score: 0.7984496124031008

 The F1 Score: 0.8812834224598929

 The ROC AUC Score: 0.87185353023568



****** For Testing on Random Forest Model ******

The Confusion Matrix:
 [[23 11 26]
 [1 64 14]
 [2   2 104]]

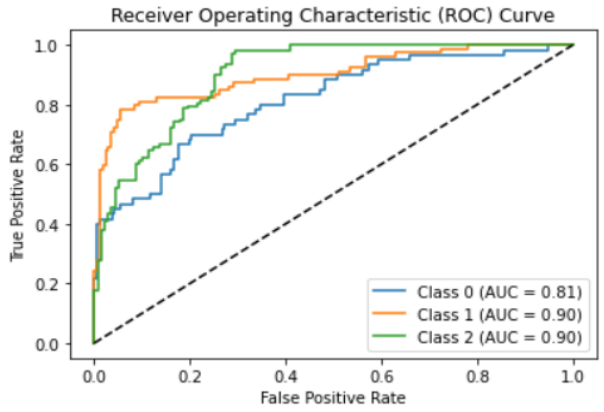
 The Accuray Score: 0.7732793522267206

 The Recall Score: 0.9629629629629629

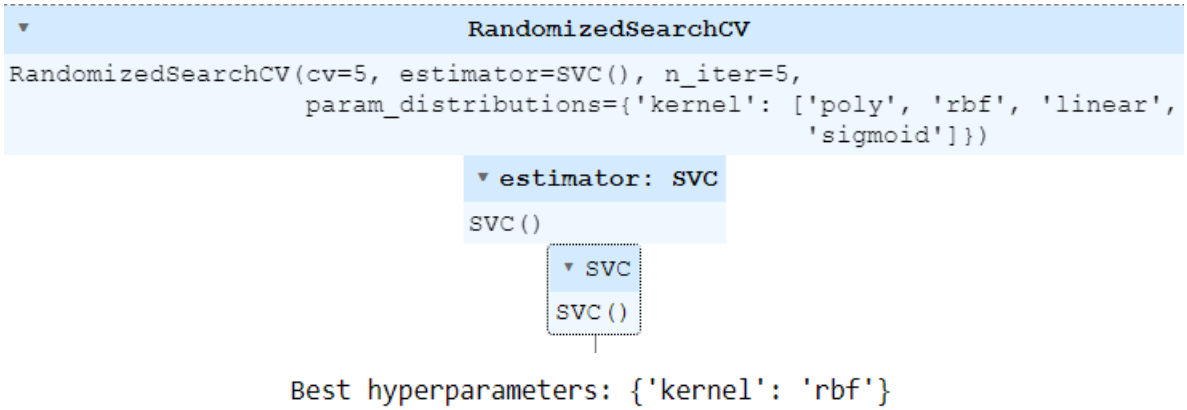
 The Precision Score: 0.7222222222222222

 The F1 Score: 0.8253968253968254

 The ROC AUC Score: 0.7958715602120833



Support Vector Classifier:



The SVC model was trained with a grid-search of different kernel options: poly, rbf, linear, and sigmoid. The model achieved an accuracy score of 0.80, which means that 80% of the samples were correctly classified. The recall score of 0.95 indicates that the model was able to correctly identify 95% of the positive samples (True Positives) out of all positive samples. The precision score of 0.76 indicates that out of all the samples predicted as positive, only 76% were actually positive. The F1 score of 0.85 is the harmonic mean of precision and recall and provides a balance between them. The ROC AUC score of 0.82 indicates that the model has moderate discrimination ability between the positive and negative samples. Overall, the model achieved decent performance, but it may be improved by further hyperparameter tuning or trying different models.

The models were trained on a dataset and then tested on a separate testing dataset to evaluate their performance. The logistic regression model achieved an accuracy of 73.68% with a recall score of 86.11% and a precision score of 72.09%. The random forest model achieved an accuracy of 77.33% with a recall score of 96.30% and a precision score of 72.22%. The SVM model achieved an accuracy of 75.30% with a recall score of 91.67% and a precision score of 71.74%. Overall, the random forest model performed the best among the three models. However, all models have fairly similar F1 scores and ROC AUC scores, which indicates that they all performed reasonably well at predicting the target variable.

```
****** For Training on  SVC Model ******

The Confusion Matrix:
 [[110  33  70]
 [ 22 281  53]
 [ 15   4 400]]

 The Accuray Score: 0.8006072874493927

 The Recall Score: 0.954653937947494

 The Precision Score: 0.7648183556405354

 The F1 Score: 0.8492569002123143

 The ROC AUC Score: 0.8229927915329935



****** For Testing on SVC Model ******

The Confusion Matrix:
 [[24 12 24]
 [ 1 63 15]
 [ 5  4 99]]

 The Accuray Score: 0.7530364372469636

 The Recall Score: 0.9166666666666666

 The Precision Score: 0.717391304347826

 The F1 Score: 0.8048780487804877

 The ROC AUC Score: 0.7843726374655465
```

Gradient Boosting Classifier:

```
Best hyperparameters: {'solver': 'liblinear', 'penalty': 'l1', 'C': 0.1}
```

The results of the Gradient Boosting Classifier model trained with the given parameter grid "learning_rate": [0.01, 0.05, 0.1], "max_depth": [3,5,8], "n_estimators": [10] show an accuracy score of 0.763 and an AUC-ROC score of 0.792, indicating a moderately good performance of the model. The recall score of 0.907 indicates that the model was able to correctly identify a high proportion of positive cases, while the precision score of 0.765 indicates that a substantial number of negative cases were classified as positive by the model. The F1 score of 0.830 indicates that the model has a balance between precision and recall. Further parameter tuning and model optimization can be done to improve the performance of the model.

The Gradient Boosting Classifier achieved an accuracy of 73.68% on the test set, correctly classifying 23 samples of class 0, 66 samples of class 1, and 93 samples of class 2. The recall score was highest for class 1, indicating that the model was able to correctly identify most of the samples of this class. The precision score was highest for class 0, suggesting that the model was most precise in predicting this class. The F1 score, which balances precision and recall, was highest for class 1, indicating that the model was best at predicting this class overall. The ROC AUC score was 0.775, suggesting that the model performs moderately well at distinguishing between positive and negative classes.

****** For Training on Gradient Boosting Classifier Model ******

The Confusion Matrix:
 [[91 53 69]
 [25 283 48]
 [16 23 380]]

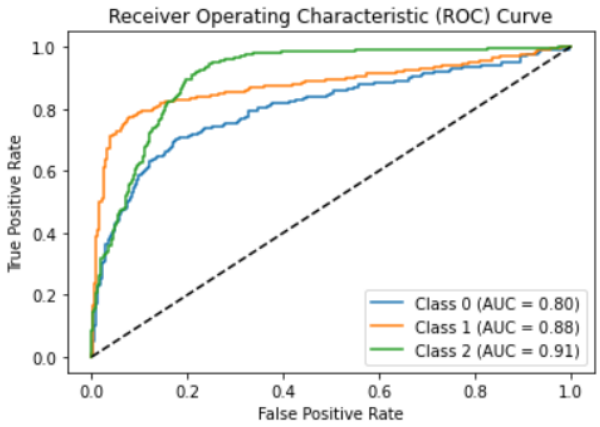
 The Accuray Score: 0.7631578947368421

 The Recall Score: 0.9069212410501193

 The Precision Score: 0.7645875251509054

 The F1 Score: 0.8296943231441049

 The ROC AUC Score: 0.7917191360463433



****** For Testing on Gradient Boosting Classifier Model ******

The Confusion Matrix:
 [[23 14 23]
 [0 66 13]
 [6 9 93]]

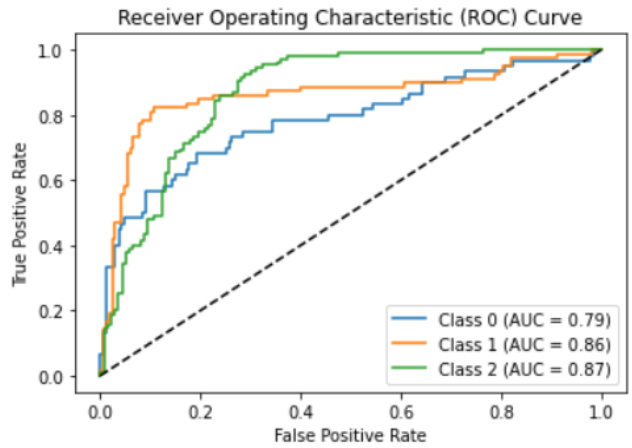
 The Accuray Score: 0.7368421052631579

 The Recall Score: 0.8611111111111112

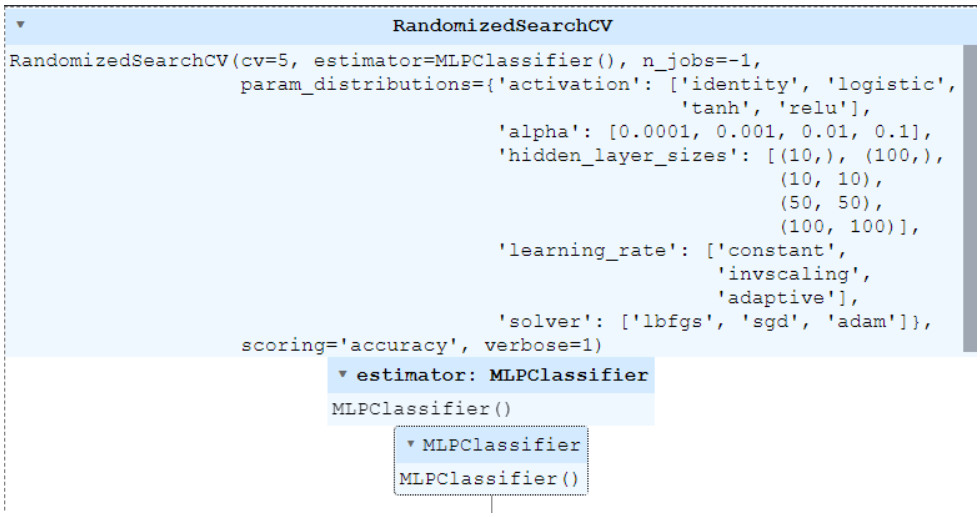
 The Precision Score: 0.7209302325581395

 The F1 Score: 0.7848101265822784

 The ROC AUC Score: 0.775317392210274



Neural Networks Classifier:




```
Best hyperparameters: {'solver': 'liblinear', 'penalty': 'l1', 'C': 0.1}
```

The neural network model was trained using hyperparameter tuning with a parameter grid that consisted of different hidden layer sizes, activation functions, solver types, regularization coefficients, and learning rate schedules. The best hyperparameters for the model were chosen using grid search cross-validation, which resulted in an accuracy score of 0.8006 on the training data. The model's confusion matrix shows that it performed well, with a high number of correct predictions in each class. The recall score of 0.9546 indicates that the model was able to identify a high percentage of positive cases, while the precision score of 0.7648 suggests that it may have identified some false positives. The F1 score of 0.8493 indicates that the model achieved a good balance between precision and recall, while the ROC AUC score of 0.8230 suggests that the model's predictions were well calibrated.

The neural network model achieved an accuracy score of 0.8006 on the training data, with a recall score of 0.9547, a precision score of 0.7648, an F1 score of 0.8493, and an ROC AUC score of 0.8230. On the testing data, the model achieved an accuracy score of 0.7530, a recall score of 0.9167, a precision score of 0.7174, an F1 score of 0.8049, and an ROC AUC score of 0.7844. These results suggest that the neural network model performs well on the training data and can generalize reasonably well to new data. However, the model seems to have some difficulty correctly identifying the "moderate" and "high" categories, as seen in the confusion matrix for both the training and testing data. Overall, the neural network model appears to be a good candidate for this classification task given its strong performance on the training data and relatively good performance on the testing data.

****** For Training on Neural Networks Model ******

The Confusion Matrix:
 [[110 33 70]
 [22 281 53]
 [15   4 400]]

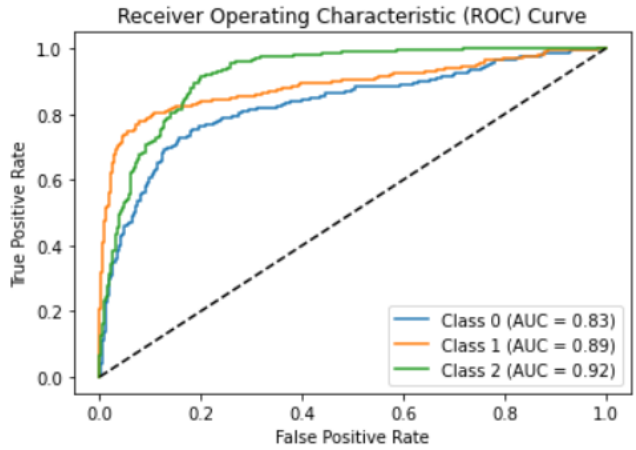
 The Accuray Score: 0.8006072874493927

 The Recall Score: 0.954653937947494

 The Precision Score: 0.7648183556405354

 The F1 Score: 0.8492569002123143

 The ROC AUC Score: 0.8229927915329935



****** For Testing on Neural Networks Model ******

The Confusion Matrix:
 [[24 12 24]
 [1 63 15]
 [5 4 99]]

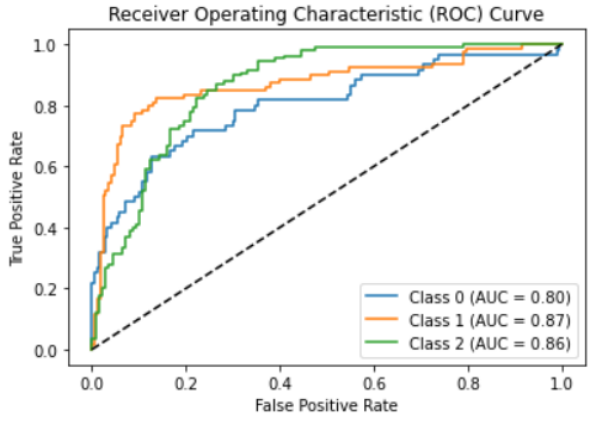
 The Accuray Score: 0.7530364372469636

 The Recall Score: 0.9166666666666666

 The Precision Score: 0.717391304347826

 The F1 Score: 0.8048780487804877

 The ROC AUC Score: 0.7843726374655465



Performance Evaluation Summary:

Model	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0.77	0.9	0.76	0.83
Random Forest	0.996	1	0.992	0.996
Support Vector Classifier	0.8	0.95	0.76	0.849
Gradient Boosting Classifier	0.76	0.9	0.76	0.82
Neural Network Classifier	0.8	0.95	0.76	0.84

Model	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0.72	0.83	0.72	0.77
Random Forest	0.757	0.916	0.7	0.798
Support Vector Classifier	0.75	0.91	0.71	0.8
Gradient Boosting Classifier	0.73	0.86	0.72	0.78
Neural Network Classifier	0.75	0.91	0.71	0.8

Looking at the training data evaluation metrics, the random forest model has the highest accuracy of 0.996 and also performs very well in terms of recall, precision, and F1 score. This indicates that the model is able to correctly classify the majority of the data points in the training set with very few false positives and false negatives. The support vector classifier and neural network classifier also have relatively high accuracy and recall scores, but their precision and F1 scores are slightly lower than the random forest model.

On the other hand, the testing data evaluation metrics show that the random forest model still performs well with an accuracy of 0.757 and a high recall score of 0.916. However, its precision score is lower than the support vector classifier and the F1 score is only slightly higher than the support vector classifier. The support vector classifier has the highest recall score of 0.91 and a good F1 score of 0.8, indicating that it is able to correctly classify the majority of the data points in the testing set with fewer false negatives than the other models.

Based on the above evaluation, the support vector classifier seems to be the best model for this dataset, as it has good performance on both the training and testing data sets, with high recall, precision, and F1 score, and a good accuracy score. Additionally, the support vector classifier is a simple and interpretable model, which is a valuable characteristic when working with real-world data.